Homework #5
406.424 Internet Applications
Spring, 2010

Due Date: May 27, 2010

In this homework, each of you will be given with a unique corpus of documents. Your first task is to extract three most frequent terms from the corpus using Galago tokenizer. Then, identify ten most associated terms for each of three terms according to the following measures:

- (25 points) Mutual information
- (25 points) Expected Mutual information
- (25 points) Chi-square
- (25 points) Dice's coefficient

You're asked to submit the Java program along with the lists of the associated terms for each of top three frequent terms.

(Note) All example Java codes and the corpus necessary for the homework tasks will be available at course home page.

***Submit your answers in MS-Word or PDF format via email to the instructor cc TA.***